研究论文

新兴技术发现模型研究*

任智军[1,2]　乔晓东[1]　张江涛[2]
[1](中国科学技术信息研究所　北京　100038)
[2](国家知识产权局中国专利信息中心　北京　100088)

摘要:【目的】在论文和专利中识别并发现待选新兴技术。**【方法】**采用 LDA 模型寻找技术主题, 使用新兴技术相似度识别待选新兴技术。利用电动汽车数据进行实验分析。**【结果】**实验结果表明, 该方法区别于以往的新兴技术识别方法, 自动识别出电动汽车领域的 25 个新兴技术。**【局限】**没有进行专家打分实验, 模型分析结果未与人工结果进行对比。**【结论】**新兴技术发现模型可高效发现新兴技术, 有效减少专家阅读文献的数量。

关键词: 新兴技术　论文　专利　电动汽车　技术相似度
分类号: G35

1　引　言

美国宾夕法尼亚大学沃顿商学院的 Day 等[1]的研究认为, 新兴技术是建立在科学基础上的革新, 它们可能创立一个新行业或改变某个老行业。它们可以是产生于突破性创新的间断性技术, 或是通过集中融合多个过去的独立研究成果而形成的更具创新性的技术。赵振元等[2]认为新兴技术是指那些新近的, 甚至正在发展的, 对经济机构产生重大影响的高技术。随着以"大众创业, 万众创新"为目标的创新时代的来临, 人们越来越重视创新技术的发展, 创新不仅要跟踪技术变化, 更要捕捉发展动态和抓住技术机会, 而技术机会分析中重要的工作之一就是找到新兴技术, 因此如何准确识别新兴技术, 进而对新兴技术进行评价和选择, 实现新兴技术的产业化具有非常重要的现实意义。

2　研究现状

现有新兴技术发现方法主要包括人工方法[3-4]、基于属性综合评价的方法[5]、基于文本挖掘方法[6-8]、基于判定规则[9-10]等方法进行新兴术语识别。

魏国平[3]利用专家打分法对新兴技术进行识别。谈毅等[4]根据新兴技术的不确定性与风险, 提出将技术路线图与实物期权方法相结合的新兴技术识别和选择框架模型。上述方法基于非系统过程, 并依靠专家的主观意见, 专家意见很容易受到专家水平和主观倾向的影响, 不同的专家给出的评分也不相同, 使评价结果的适用性大打折扣。

黄鲁成等[5]在属性集和属性测度理论基础上提出属性综合评价和决策系统, 对一组技术进行判别, 进而对其进行分类, 以求找出其中的新兴技术, 然而在部分指标的打分中, 依然使用的是专家打分, 然后利用 AHP 得出技术权重, 如果待识别的技术过多就会造成识别速度过慢, 同时专家主观性较强, 评价结果依然无法被广泛使用。

Kostoff 等[6]使用文本挖掘的方法对断裂性技术进行识别。王凌燕等[7]利用专利文献提出识别新兴技术主题的初步技术框架, 并以工业生物技术领域的专利文献为数据来源, 采用文本聚类技术、共词战略坐标分析、共词网络分析、专利分析等方法对该技术领域的新兴技术主题进行实证分析。在海量文本中找到待评价技术的文本挖掘方法主要依靠词聚类, 结果形成技术方向, 而不是技术主题, 同时技术方向中技术术语是排他的, 而不是兼容的, 割裂了技术之间的关系, 因此文本挖掘的方法也不能较好地发现新兴技术, 其

通讯作者: 任智军, ORCID: 0000-0002-3269-3556, E-mail: renzhijun@cnpat.com.cn。
*本文系中国博士后科学基金资助项目"基于论文与专利整合的分析与挖掘方法研究"(项目编号: 2013M540125)的研究成果之一。

主要方法也是对指定技术进行是否为新兴技术的判定。李蓓等[8]采用基于专利引用耦合聚类的新兴技术识别模型及其相关指标体系，并以美国专利商标局(简称美专局)授权的专利数据库为数据源，对纳米技术领域展开了实证分析。与文献[7]的结果一致，研究的是技术方向不是技术主题，同时聚类的主题过于粗糙，技术识别较少。

Kim 等[9]提出基于判定规则的方法，采用先验知识进行验证，提出判定规则，最后进行检验分析。在笔者的前期研究[10]中，也采用了基于判定规则的方法，分别是顶尖热点技术、萌芽新兴技术、趋稳新兴技术、成长新兴技术和衍生新兴技术 5 种新兴技术，每种新兴技术给出了计算指标，然而这种基于单一指标判定出技术术语的方法仍然是基于规则的，是一种经过试错产生的办法，无法进一步改进，不能够利用既有数据进行有效学习。

在分析现有研究的过程中发现，专家评价在得到足够的支持信息后识别新兴技术的方法已经成熟，其他研究的主要目的在于提供各种方法支持，减少专家的工作强度，提高信息支持。因此笔者认为发现新兴技术的关键问题如下:

(1) 如何从文献中直接找到所有可能的新兴技术;

(2) 减少待分析新兴技术的数量，提高专家工作效率;

(3) 新兴技术的专家判定方法如何得到足够的信息支持，同时减少专家阅读文献的数量和减少专家的主观性。

针对以上问题，本文采用主题模型，从海量的科技文本(论文和专利)中找到待评价的新兴技术主题;在数据整合的基础上，使用新兴技术相似度算法，对待评价新兴技术排序，并根据阈值进行新兴技术截取，进而自动找到候选新兴技术，降低待分析新兴技术的数量;标注新兴技术，并提供新兴技术的核心文献，为专家判定提供信息支持。

3　研究思路与框架

新兴技术识别的重要工作是在海量文本中自动发现新兴技术候选集，并准确识别候选集，找到最可能是新兴技术的技术主题。因此新兴技术发现模型包括论文与专利共同研究技术主题模型、新兴技术识别模型和新兴技术选择和标注三个组成部分: 共同研究技术主题模型首先对论文和专利进行整合，在整合数据的基础上使用主题模型发现领域所有相关的技术主题，然后采用二元判定找到文献与主题的关系，形成技术主题集合; 新兴技术识别模型利用现有新兴技术数据，根据特征发现新兴技术发展的技术趋势和技术内涵，然后根据技术主题和新兴技术的相似关系进行相似度计算，给出相似度排序; 新兴技术选择和标注主要根据相似度排序缩小新兴技术集合，然后在小集合中进行新兴技术标注，最后将待判定集合交予专家做最后判定，如果是新兴技术便入库，否则抛弃。新兴技术发现模型框架如图 1 所示:

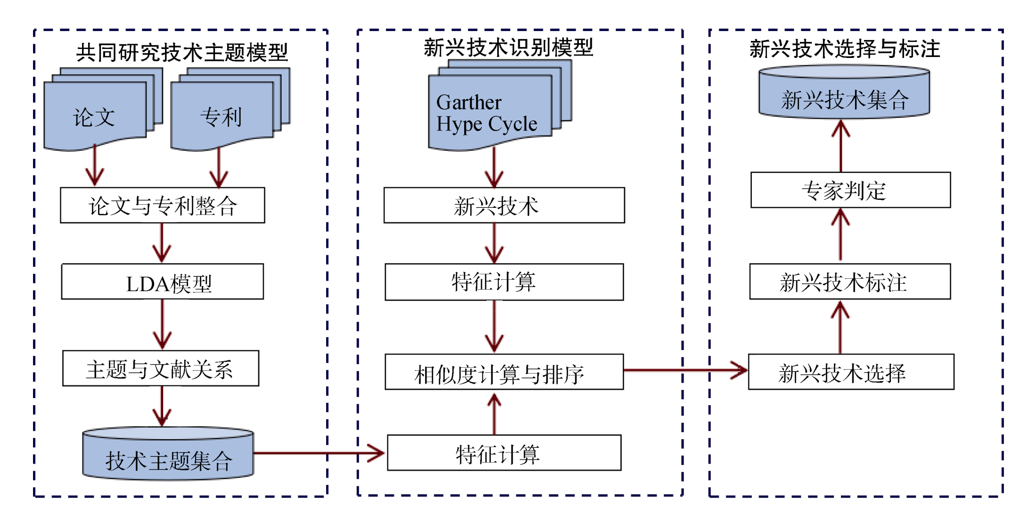


图 1　新兴技术发现模型框架

4 实验过程

4.1 数据收集与整理

新兴技术发现模型中论文与专利整合的目的在于论文是基础研究成果的表现形式，专利文献是技术创新成果的表现形式，找到研究成果与技术创新成果中全部的技术主题，才能够发现现有新兴技术中全部的科技创新技术，识别本领域真实的新兴技术。

论文和专利整合首先是数据收集整理，要对中国专利著录项目和中国论文期刊的摘要等数据进行整理。之后整合论文(专利)标题、论文(专利)摘要、论文发表年份(专利公开年份)、论文机构(专利申请人)、论文作者(专利发明人)等信息，整合的数据如图 2 所示：

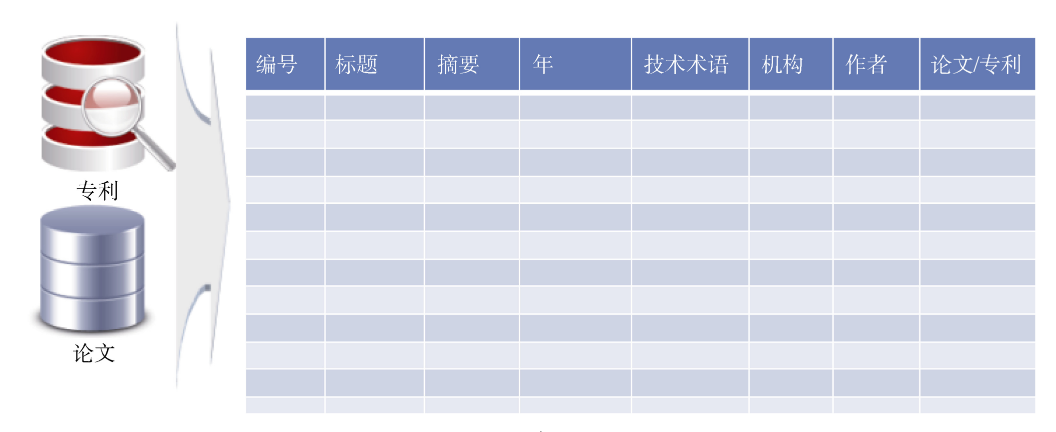


图 2 论文与专利整合结果

为验证基于论文与专利整合的新兴技术发现模型的可行性，选择电动汽车的论文和专利数据进行实例研究，从万方数据库和中国国家知识产权局专利检索系统 CPRS 中获取 2009 年–2013 年数据，包括中国电动汽车论文 31 258 篇，中国电动汽车专利 104 291 篇，共计 135 549 条数据，用以构建电动汽车论文和专利整合数据库。将提取出的论文和专利的关键词加入到分词词库中，对论文和专利进行分词，去除噪音词后，共得到 243 756 个关键词。

4.2 技术主题模型

技术主题是由具有强文本表示功能的特征关键词组成，而强文本表示功能是指在文本表示时，能将文本的内容及特征(例如领域类别、主题思想、中心意义等)鲜明地表示出来[11]。主题模型是统计学中用以识别文字中隐含主题的一种建模方法[12]，因此本文采用 LDA(Latent Dirichlet Allocation)主题模型作为新兴技术主题的计算模型。笔者采用 Gibbs 采样方法进行 LDA 模型建模，得到文档–主题矩阵和主题–词矩阵，其中主题–词概率矩阵即为本文的技术主题，文档–主题矩阵是技术主题和文献之间概率关系。

LDA 主题模型形成了文献和主题之间的概率关系，为了进一步识别新兴技术，需要将概率关系转化为 0 和 1 的二元关系，即找到表达文献主旨的技术主题。为准确表达技术主题和文献之间的关系，笔者设定一个固定的阈值，大于阈值设为 1，小于阈值设为 0，即只有达到这个阈值的主题才能作为这篇文献的标引主题，技术主题和文献关系确定如图 3 所示：

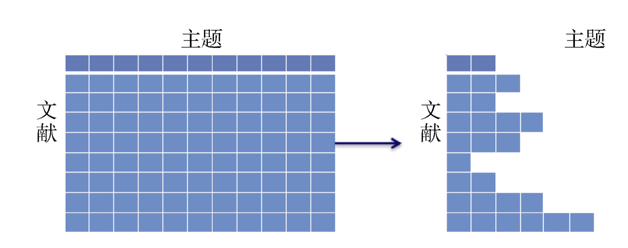


图 3 技术主题和文献关系

利用整合后的数据库，共识别 243 756 个关键词。针对电动汽车论文和专利的整合数据，利用 LDA 算法进行技术主题识别，参数设置为 800 个主题，$\alpha=0.1$，$\beta=0.01$，进行 1 000 次迭代。

4.3 新兴技术识别方法

在论文与专利整合数据中，利用主题模型获取技术主题和文献与技术主题强关系，在此基础上，需要找到一批已知的新兴技术作为参照，利用 LDA 识别的技术主题与已知的新兴技术进行相似度计算，相似度较大则说明此技术主题可能为新兴技术，相似度较小

说明此技术主题成为新兴技术的可能性较小，或者说历史上这种模式的新兴技术较少或没有。因此，新兴技术识别模型就是要找到一批新兴技术，并把识别出的技术主题与这些新兴技术进行相似度计算。研究分析机构 Gartner 每年都会推出不少研究报告，不过其中最受关注的是新兴技术成熟度曲线(也称为新兴技术炒作周期报告)，因此，笔者利用 Hype Cycle for Emerging Technologies 中的技术作为新兴技术训练集，而后采用相似度算法，计算识别出的技术主题和 Gartner 的新兴技术数据之间的相似度，并根据相似度进行排序，最终得到一个技术主题有序队列，由专家选择判断。

识别新兴技术主要分为学习和预测两个阶段：学习阶段，收集整理 Gartner 的新兴技术数据，计算新兴技术的技术特征；预测阶段，对技术主题利用技术主题和文献关系计算出技术特征，利用技术特征和相关度算法计算技术主题的相关度，并排序。具体如图 4 所示：

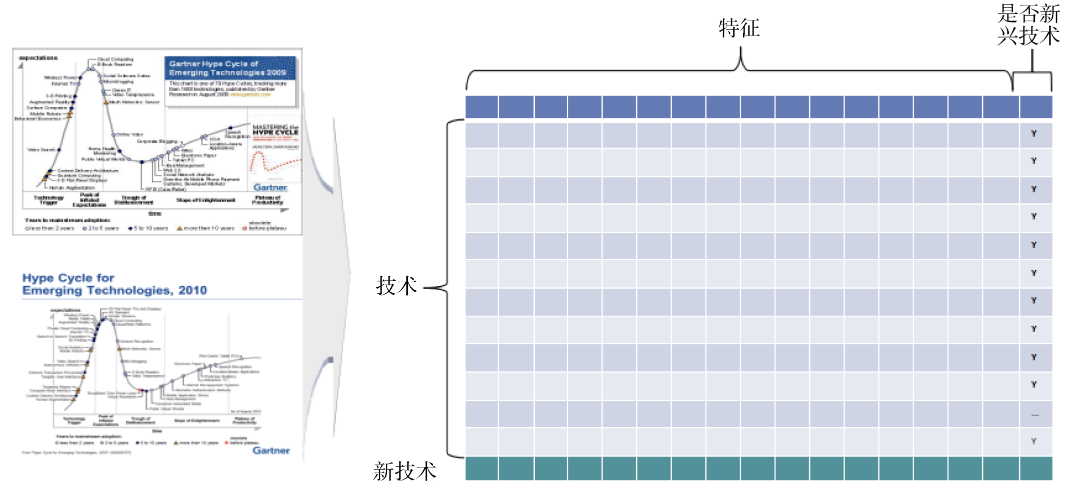


图 4　新兴技术识别模型学习及预测图

4.4　新兴技术相似度算法

新兴技术识别的核心是排序，通过技术主题与 Gartner 技术成熟度曲线中的新兴技术的相似程度进行判定。技术相似度越大，被选择的可能性越大，则识别为新兴技术的可能性越大。为了计算技术主题和新兴技术的相似程度，笔者采用余弦相似度算法计算技术主题和新兴技术的相似度，公式如下：

$$\mathrm{Sim}(T_i,E_j)=\frac{\sum_{k=1}^{n}F_{ik}\times F_{jk}}{\sqrt{(\sum_{k=1}^{n}F_{ik}^2)(\sum_{k=1}^{n}F_{jk}^2)}} \quad (1)$$

其中，T 是技术主题，E 是 Gartner 的新兴技术，F 是特征属性。在得到技术主题和单个新兴技术相似度后，从中找出最相似的 M 个新兴技术，累加技术主题与 M 个新兴技术的相似度并求平均，公式如下：

$$\mathrm{ATS}_i=\frac{\sum_{j=0}^{M}\mathrm{Sim}(T_i,E_j)}{M} \quad (2)$$

新兴技术的相似度算法如下：

输入：训练好的 Gartner 新兴技术及特征向量，技术主题及特征向量，M

输出：技术主题，技术主题和新兴技术相似度列表

①根据特征项集合重新描述训练科技术语及特征向量，并构建 KDTree;

②找出 M 个最相似的新兴技术;

③计算新科技术语与 M 个新兴技术的相似度;

④将 M 个相似度加和求平均，并记录到 Hash 表中，其中 Key 是技术术语，Value 是技术相似度;

⑤将 Hash 表根据 Value 从大到小排序，并输出。

Gartner 认为新兴技术是一些受到炒作而成为关注焦点的技术，或者是有可能带来重大影响的技术。Gartner 技术成熟度曲线的绘制指标主要有参加会议人数[13-14]、技术创新的文献数量或比例[15-18]、专利统计数据[19-20]。从 Gartner 对新兴技术的定义和对新兴技术绘制的技术成熟度曲线中可以看出，新兴技术应受到炒作或者关注，即需要有研发机构和研究人员对此技术进行研

究和工程化(机构研发系数、作者研发系数)。论文发表数量和专利申请数量所达到的频度(论文频度和专利频度)，技术截止到目前已经出现的时间(年度特征)对技术是否判定为新兴技术都有影响，同时技术归属度、领域相关度等指标能够反映技术与领域的相互关系，因此笔者将以上指标称之为内容性特征。目前，很多学者提出以趋势性指标[9-10](增长率、相对增长率、作者占有率、作者增长率、机构占有率、机构增长率)绘制 Hype Cycle。本文结合趋势性指标和内容性指标，通过更多角度和维度，识别新兴技术。指标内容如表 1 所示:

表 1　新兴技术指标集

指标类型	指标名称		指标内容
内容性指标	频度	论文	某个技术主题在年度论文中超过一定阈值的时候，该频度为 1，否则为 0。这个阈值是变化的，一般采用采集的 Gartner 新兴技术中的最小值，本次实验论文阈值是 2。
		专利	某个技术主题在年度专利中超过一定阈值的时候，该频度为 1，否则为 0。这个阈值是变化的，一般采用采集的 Gartner 新兴技术中的最小值，本次实验专利阈值是 1。
	年数	论文	某个技术主题在论文中出现总年数。
		专利	某个技术主题在专利中出现总年数。
	归属度	论文	某个技术主题在本领域中出现的频率与在总的论文库中出现频率的比率。
		专利	某个技术主题在本领域中出现的频率与在总的专利库中出现频率的比率。
	领域相关度	论文	某个技术主题与领域的互信息。
		专利	某个技术主题与领域的互信息。
	机构研发系数	论文	论文机构研究系数：论文数与机构数量的比值。
		专利	申请人研发系数：专利申请数与申请人数量的比值。
	作者研发系数	论文	论文数与作者数量的比值。
		专利	专利数与发明人数量的比值。
趋势性指标	增长率	论文	(累积到本年的论文发表量–上一年累积发表量)/累积到上一年的论文发表量
		专利	(累积到本年的专利申请量–上一年累积专利申请量)/累积到上一年的专利申请量
	相对增长率	论文	(本年的论文发表量–上一年发表量)/上一年的论文发表量
		专利	(本年的专利申请量–上一年专利申请量)/上一年的专利申请量
	作者占有率	论文	本年作者数量/本年总的作者数量
		专利	本年发明人数量/本年总的发明人数量
	作者增长率	论文	(本年的作者数量–上一年作者数量)/上一年作者数量
		专利	(本年的发明人数量–上一年发明人数量)/上一年发明人数量
	机构占有率	论文	本年研究机构数量/本年总的研究机构数量
		专利	本年研究申请人数量/本年总的申请人数量
	机构增长率	论文	(本年的研究机构数量–上一年研究机构数量)/上一年研究机构数量
		专利	(本年的申请人数量–上一年申请人数量)/上一年申请人数量

4.5　新兴技术选择与标注

在计算得到技术主题与 Gartner 新兴技术的相似度以及相似度排序后，设定一个阈值，选择有限个技术主题(即待选新兴技术)交给专家进行判定。

在选择得到待选新兴技术主题后，使用一个词或者一个词组对新兴技术名称进行标注，由于主题模型的标注比较复杂，在本模型中主要采用人工标注的方式进行。

进行技术主题标注后，在交予专家判定前，需要给专家提供待选新兴技术的支持信息，即提供待选新兴技术的关键文档，本文选择 LDA 模型的文档–主题矩阵中，与技术主题最相关的前 5 篇论文和前 5 篇专利提供给专家。专家经过研读后，如果判定为新兴技术，则将此技术入库，否则不予入库。

5　实验结果

5.1　论文与专利技术主题模型实验结果

根据 4.2 节的实验，笔者共获取 1 000 个主题，其中前 20 个主题的部分结果如表 2 所示。

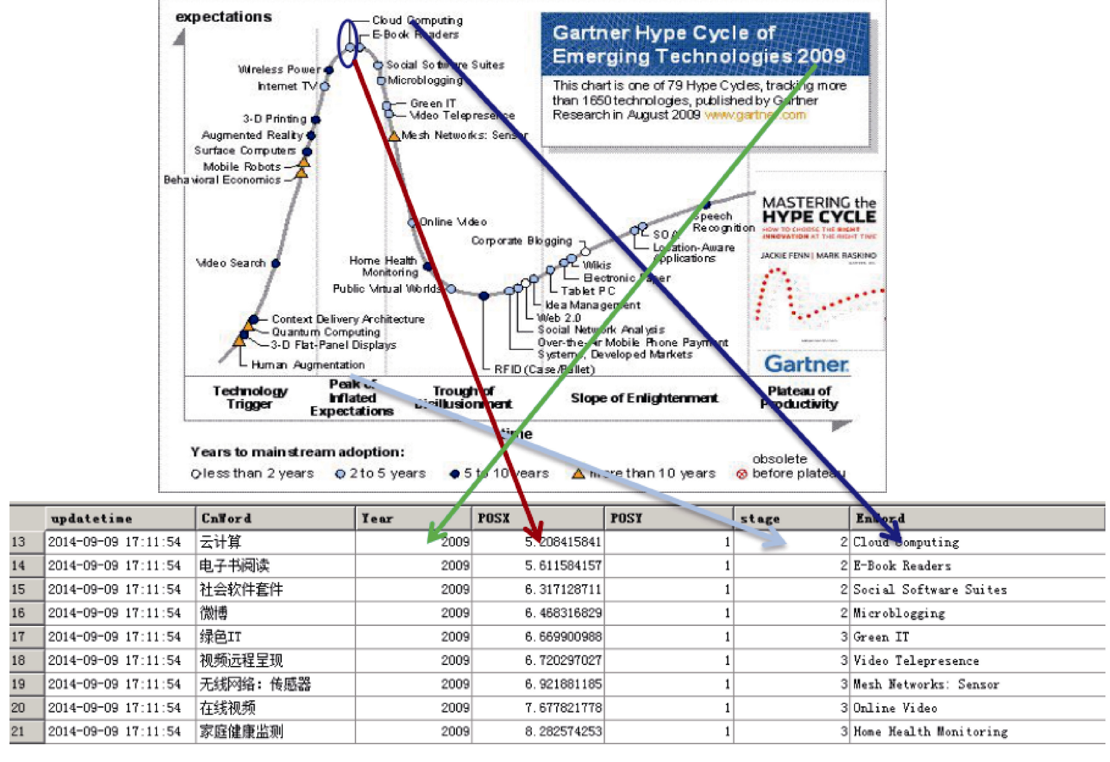
表 2　电动汽车主题模型部分结果

主题 1		主题 2		主题 3		主题 4	
词	权重	词	权重	词	权重	词	权重
驱动电源	0.074460682	线圈骨架	0.030627871	密封构件	0.032125206	防撞杆	0.101145038
超声电机	0.030619346	橡胶圈	0.030627871	开口槽	0.026359143	燃料电池组	0.026717557
调整装置	0.016701461	弹性片	0.022970904	充电组件	0.022240527	排出口	0.025763359
收集器	0.016005567	导电线	0.016845329	座椅主体	0.019769357	电导体	0.02480916
驱动器	0.016005567	车载设备	0.016079632	支座本体	0.016474465	腰带头	0.016221374
主题 5		主题 6		主题 7		主题 8	
词	权重	词	权重	词	权重	词	权重
粉末冶金	0.032467532	保险杠	0.05849359	催化剂层	0.049905482	充电插座	0.068493151
滑动门	0.027829314	主框架	0.03125	燃料电池	0.022306238	充电插头	0.0498132
连接端	0.019480519	前围板	0.025641026	电解质膜	0.020415879	点火开关	0.04109589
温控器	0.017625232	前端部	0.018429487	气体扩散层	0.017013233	恒流二极管	0.02615193
冷却箱	0.016697588	加强件	0.013621795	高分子电解质	0.016824197	电磁开关	0.02532171

同时计算技术主题和文献关系，设置阈值大于 0.0125 以上的标记为 1，否则为 0。

5.2　新兴技术识别模型实验结果

为了验证新兴技术识别模型的有效性，笔者收集了 Gartner 的 Hype Cycle 图，根据图中数据采集文字、技术周期图、发表时间、位置等信息(如图 5 所示)，将对应的英文翻译成中文，同时去除部分概念数据，比如 Disaster Recovery Service-Level Management(灾难恢复服务水平管理)等，共采集有效可以被用作新兴技术学习数据 650 个。Gartner 新兴技术指标计算是根据翻译得到的中文新兴技术术语和年份进行检索，检索得到结果后，再根据特征公式进行计算，得到特征值。



	updatetime	CnWord	Year	POSX	POSY	stage	EnWord
13	2014-09-09 17:11:54	云计算	2009	5.208415841	1	2	Cloud Computing
14	2014-09-09 17:11:54	电子书阅读	2009	5.611584157	1	2	E-Book Readers
15	2014-09-09 17:11:54	社会软件套件	2009	6.317128711	1	2	Social Software Suites
16	2014-09-09 17:11:54	微博	2009	6.468316829	1	2	Microblogging
17	2014-09-09 17:11:54	绿色IT	2009	6.669900988	1	3	Green IT
18	2014-09-09 17:11:54	视频远程呈现	2009	6.720297027	1	3	Video Telepresence
19	2014-09-09 17:11:54	无线网络：传感器	2009	6.921881185	1	3	Mesh Networks: Sensor
20	2014-09-09 17:11:54	在线视频	2009	7.677821778	1	3	Online Video
21	2014-09-09 17:11:54	家庭健康监测	2009	8.282574253	1	3	Home Health Monitoring

图 5　Gartner 技术数据库存储样式

每个新兴技术利用 CPRS(词和公开年)和百度学术(词和发表年)进行检索，对检索结果进行统计得出 4.4 节中提出的指标。

根据技术主题模型中获取的 1 000 个主题和每个技术主题 24 个特征，M 取 50，利用公式(2)和公式(3)计算新兴技术相似度，再根据新兴技术相似度进行排序，将相似度最高的前 25 个技术主题作为待选新兴技术，如表 3 所示：

表 3　相似度最高的 25 个技术主题表

编号	技术主题	关键词	相似度
1	386	充电设施；充换电服务；标准体系；商业模式；电磁兼容	0.8714
2	172	燃料电池堆；电池堆；燃料电池；燃料电池系统；阳极侧	0.8591
3	194	蓄电池；蓄电池组；快速充电；充电回路；充电电流	0.8551
4	235	直流电动机；直流电机；换向器；直流发电机；驱动单元	0.8520
5	149	充电电池；电池充电；充电模式；充电器；充电装置	0.8516
6	51	太阳能电池；染料敏化；工作电极；太阳电池；光电转换效率	0.8486
7	528	混合动力汽车；混合动力；发动机；控制策略；混合动力系统	0.8462
8	501	燃料电池；燃料电池系统；燃料电池堆；单电池；燃料气体	0.8461
9	426	质子交换膜；膜电极；催化层；燃料电池；聚合物电解质	0.8434
10	77	开关磁阻电机；磁阻电机；开关磁阻；功率变换器；调速系统	0.8404
11	324	控制系统；无刷直流电机；单片机；直流电机；实验结果	0.8397
12	442	碳纤维；汽车车身；主表面；空心管；折叠部	0.8397
13	265	充电桩；充电机；充电站；监控系统；充电设备	0.8396
14	690	控制模块；无线接收模块；显示屏；无线发射模块；驾驶员	0.8377
15	31	锂离子电池；负极片；正极片；电池芯；正极集流体	0.8368
16	412	动力总成；发动机；悬置系统；发动机悬置；悬置支架	0.8348
17	64	驱动器；转向电机；智能控制器；测试电路；电动机	0.8319
18	498	电动助力转向系统；控制策略；操纵稳定性；横摆角速度；仿真结果	0.8294
19	780	储能蓄电池；智能充放电；电池组件；锂离子电池；太阳能	0.8284
20	193	再生制动；能量回收；制动能量；制动能量回收；制动系统	0.8235
21	208	电机控制器；驱动电机；整车控制器；控制系统；动力电池组	0.8216
22	195	电池单元；电池系统；电池管理系统；充电状态；电池模块	0.8018
23	453	永磁同步电机；同步磁阻电机；异步电机；悬浮力；无轴承	0.7934
24	168	石墨烯；复合材料；二氧化钛；水溶液；混合溶液	0.7700
25	266	无线充电；接收器；发射线圈；接收线圈；无线充电器	0.7286

5.3　新兴技术标注与信息支持实验结果

在获取待选新兴技术后，标注待选新兴技术和提供相关论文和专利，根据阅读主题中的关键词和部分最相似文献，确定技术主题的标注信息，具体信息如表 4 所示。

可见，在电动汽车论文和专利数据中，选出待选新兴技术分别是“充换电服务”、“燃料电池堆”、“快速充电”、“直流电动机”、“充电电池”、“太阳能电池”、“混合动力汽车”、“燃料电池”、“质子交换膜”、“开关磁阻电机”、“控制系统”、“碳纤维”、“充电监控”、“无线移动”、“锂离子电池”、“发动机悬置”、“智能控制”、“电动助力转向”、“智能充放电”、“制动能量回收”、“整车控制器”、“电池管理系统”、“永磁同步电机”、“石墨烯”、“无线充电”等 25 个技术。

然后将电动汽车 LDA 模型的文档–主题矩阵中与技术主题最相关的前 5 篇论文和前 5 篇专利提供给专家，限于篇幅，仅提供两个主题，如表 5 所示。

表 4　待选新兴技术主题标注表

编号	技术主题	关键词	标注信息
1	386	充电设施; 充换电服务; 标准体系; 商业模式; 电磁兼容	充换电服务
2	172	燃料电池堆; 电池堆; 燃料电池; 燃料电池系统; 阳极侧	燃料电池堆
3	194	蓄电池; 蓄电池组; 快速充电; 充电回路; 充电电流	快速充电
4	235	直流电动机; 直流电机; 换向器; 直流发电机; 驱动单元	直流电动机
5	149	充电电池; 电池充电; 充电模式; 充电器; 充电装置	充电电池
6	51	太阳能电池; 染料敏化; 工作电极; 太阳电池; 光电转换效率	太阳能电池
7	528	混合动力汽车; 混合动力; 发动机; 控制策略; 混合动力系统	混合动力汽车
8	501	燃料电池; 燃料电池系统; 燃料电池堆; 单电池; 燃料气体	燃料电池
9	426	质子交换膜; 膜电极; 催化层; 燃料电池; 聚合物电解质	质子交换膜
10	77	开关磁阻电机; 磁阻电机; 开关磁阻; 功率变换器; 调速系统	开关磁阻电机
11	324	控制系统; 无刷直流电机; 单片机; 直流电机; 实验结果	控制系统
12	442	碳纤维; 汽车车身; 主表面; 空心管; 折叠部	碳纤维
13	265	充电桩; 充电机; 充电站; 监控系统; 充电设备	充电监控
14	690	控制模块; 无线接收模块; 显示屏; 无线发射模块; 驾驶员	无线移动
15	31	锂离子电池; 负极片; 正极片; 电池芯; 正极集流体	锂离子电池
16	412	动力总成; 发动机; 悬置系统; 发动机悬置; 悬置支架	发动机悬置
17	64	驱动器; 转向电机; 智能控制器; 测试电路; 电动机	智能控制
18	498	电动助力转向系统; 控制策略; 操纵稳定性; 横摆角速度; 仿真结果	电动助力转向
19	780	储能蓄电池; 智能充放电; 电池组件; 锂离子电池; 太阳能	智能充放电
20	193	再生制动; 能量回收; 制动能量; 制动能量回收; 制动系统	制动能量回收
21	208	电机控制器; 驱动电机; 整车控制器; 控制系统; 动力电池组	整车控制器
22	195	电池单元; 电池系统; 电池管理系统; 充电状态; 电池模块	电池管理系统
23	453	永磁同步电机; 同步磁阻电机; 异步电机; 悬浮力; 无轴承	永磁同步电机
24	168	石墨烯; 复合材料; 二氧化钛; 水溶液; 混合溶液	石墨烯
25	266	无线充电; 接收器; 发射线圈; 接收线圈; 无线充电器	无线充电

表 5　提供给专家的资源列表(样例)

编号	技术主题	新兴技术标注	相关论文	相关专利
1	386	充换电服务	(1) 电动汽车充换电服务网络运营管理系统的研究与设计. 张海龙, 冯森, 李建祥.《陕西电力》. 2011 (2) 基于物联网的电动汽车智能充换电服务网络电池管理. 薛飞, 雷宪章, 张野飚. 《电力系统自动化》. 2012 (3) 电动汽车智能充换电服务网络建设与运营. 贾俊国.《电力需求侧管理》. 2011 (4) 基于复杂网络的电动汽车智能充换电服务网络评估方法. 薛飞, 雷宪章, 张野飚.《电网技术》.2012 (5) “电动汽车-车联网”商业模式研究. 叶瑞克, 陈秀妙, 朱方思宇.《北京理工大学学报: 社会科学版》. 2012	(1) 申请号: 201210042946.5 基于能量等效的电动汽车充电设施负荷预测系统及方法 (2) 申请号: 201310034395.2 电动汽车充换电服务网络发展演算仿真系统 (3) 申请号: 201110122776.7 燃料供应系统 (4) 申请号: 201210042247.0 电动汽车充电设施负荷预测系统及预测方法 (5) 申请号: 201310682498.X 基于生态模拟的电动汽车充换电服务网络仿真系统及方法
2	172	燃料电池堆	(1) 高功率薄型金属双极板 PEM 燃料电池堆研究. 王东, 王涛, 张伟.《Chinese Journal of Power Sources》. 2009 (2) 可测正负压的燃料电池单片电压检测系统设计. 邓坚, 王姣, 邓超. 《武汉理工大学学报: 信息与管理工程版》 (3) 5 kW 质子交换膜燃料电池堆之气体与水管理系统. 马小康, 郑为阳, 方富民.《武汉大学学报: 工学版》. 2012 (4) 加压燃料电池发动机开发.赵景辉, 陈沛, 吴兵.《上海汽车》. 2009 (5) Ener1 向现代重工电动汽车供应锂电池堆.《新材料产业》. 2010	(1) 申请号: 201010107129.4 燃料电池系统及判定燃料电池系统内的阳极压力传感器是否正常运行的方法 (2) 申请号: 201110176379.8 无相对湿度感测设备反馈的堆阴极入口相对湿度控制 (3) 申请号: 200710044739.2 一种控制燃料电池空气和氢气运行压力稳定的方法 (4) 申请号: 201380057771.8 燃料电池堆和用于组装该燃料电池堆的方法 (5) 申请号: 201380065084.0 燃料电池堆及使用了该燃料电池堆的载荷分担方法

6 结　语

本文提出一种基于论文与专利整合的新兴技术发现模型的研究方法，构建基于论文与专利整合分析与挖掘的知识挖掘模型，该模型融合内容性指标和趋势性指标，采用LDA和新兴技术相似度进行新兴技术识别，并通过电动汽车的论文数据和专利数据进行实例验证。实验结果表明这种方法区别于以往的新兴技术识别方法，减少了专家的主观性和待分析新兴技术的数量，提高了专家的工作效率。

但本文没有进行专家打分实验，模型分析结果未与人工结果进行对比，可能存在误差。下一步工作将进行专家打分实验，根据不一致数据进一步修改模型和方法，同时考虑到不同的特征对识别结果的影响，进行特征选择，以优化新兴技术识别模型的识别的准确性。

参考文献：

[1] 乔治·戴，保罗·休梅克. 沃顿论新兴技术管理[M]. 石莹等译. 北京：华夏出版社，2002. (Day G S, Schoemaker P. Towards Knowledge Services: Wharton on Mananging Emerging Technologies [M]. Translated by Shi Ying, et al. Beijing: Huaxia Publishing House, 2002.)

[2] 赵振元，银路，成红. 新兴技术对传统管理的挑战和特殊市场开拓的思路[J]. 中国软科学，2004 (7): 72-77. (Zhao Zhenyuan, Yin Lu, Cheng Hong. The Challenge of Emerging Technology on Traditional Management and the Thought of Developping Special Market [J]. China Soft Science, 2004(7): 72-77.)

[3] 魏国平. 新兴技术管理策略研究—基于新兴技术特征的分类分析[D]. 杭州：浙江大学, 2006. (Wei Guoping. Study on Emerging Technology Management Strategy – Classification Analysis Based on Features of Emerging Technology [D]. Hangzhou: Zhejiang University, 2006.)

[4] 谈毅，黄燕丽. 基于过程的新兴技术规划与选择模型研究[J]. 科技管理研究, 2007, 27 (8): 5-8. (Tan Yi, Huang Yanli. Study on the New Emerging Technology Planning and Choosing Model [J]. Science and Technology Management Research, 2007, 27(8): 5-8.)

[5] 黄鲁成，卢文光. 基于属性综合评价系统的新兴技术识别研究[J]. 科研管理，2009, 30(7): 190-194. (Huang Lucheng, Lu Wenguang. Study on the Identification of Emerging Technology by an Attribute Synthetic Measure Model [J]. Science Research Management, 2009, 30(7): 190-194.)

[6] Kostoff R N, Boylan B, Simons G R. Disruptive Technology Roadmaps [J]. Technological Forecasting & Social Change, 2004, 71(1): 141-159.

[7] 王凌燕，方曙，季培培. 利用专利文献识别新兴技术主题的技术框架研究[J]. 图书情报工作，2011，55(18): 74-78. (Wang Lingyan, Fang Shu, Ji Peipei. Using Patent Documents to Study the Technology Framework of Detecting Emerging Technology Topics [J]. Library and Information Service, 2011, 55(18): 74-78.)

[8] 李蓓，陈向东. 基于专利引用耦合聚类的纳米领域新兴技术识别[J]. 情报杂志，2015，34(5): 35-40. (Li Bei, Chen Xiangdong. Identification of Emerging Technologies in Nanotechnology Based on Citing Coupling Clustering of Patents [J]. Journal of Intelligence, 2015, 34(5): 35-40.)

[9] Kim J, Hwang M, Jeong D H, et al. Technology Trends Analysis and Forecasting Application Based on Decision Tree and Statistical Feature Analysis [J]. Expert Systems with Applications, 2012, 39(16): 12618-12625.

[10] 任智军，乔晓东，徐硕，等. 基于数据挖掘的技术机会发现模型研究[J]. 情报杂志，2015，34(6): 174-177. (Ren Zhijun, Qiao Xiaodong, Xu Shuo, et al. An Approach for Technology Opportunities Discovery Model Based on Data Mining [J]. Journal of Intelligence, 2015, 34(6): 174-177.)

[11] 殷蜀梅. 判断新兴研究趋势的技术框架研究[J]. 图书情报知识，2008(3): 76-80. (Yin Shumei. A Study on the Technology Framework for Detecting Emerging Trend in Medical Research [J]. Document, Informaiton & Knowledge, 2008(3): 76-80.)

[12] Blei D M, Ng A Y, Jordan M I, et al. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[13] 郭道劝. 基于 TRL 的技术成熟度模型及评估研究[D]. 长沙：国防科学技术大学, 2010. (Guo Daoquan. The Study of Technology Maturity Model and Assessment Based on TRL [D]. Changsha: National University of Defense Technology, 2010.)

[14] 杨良选. 技术成熟度多维评估模型研究[D]. 长沙：国防科学技术大学, 2011. (Yang Liangxuan. Research on Technology Maturity Multi-Dimensional Assessment Model [D]. Changsha: National University of Defense Technology, 2011.)

[15] Jarvenpaa H, Makinen S J. An Empirical Study of the Existence of the Hype Cycle: A Case of DVD Technology [C]. In: Proceedings of 2008 IEEE International Engineering

Management Conference. 2008.
[16] Lind J. Convergence: History of Term Usage and Lessons for Firm Strategists [A]. // The Anticipation of Converging Industries [M]. Springer London, 2004.
[17] Tammen H, Budde A P, Zucht H. Peptidomics Analysis of Human Blood Specimens for Biomarker Discovery [J]. Environmental Innovation and Societal Transitions, 2007, 7(5): 605-613.
[18] Jarvenpaa H, Makinen S J. Empirically Detecting the Hype Cycle with the Life Cycle Indicators: An Exploratory Analysis of Three Technologies [C]. In: Proceedings of the 2008 International Conference on Industrial Engineering and Engineering Management. 2008.
[19] Budde B, Alkemadeb F, Hekkert M. On the Relation Between Communication and Innovation Activities: A Comparison of Hybrid Electric and Fuel Cell Vehicles [J]. Environmental Innovation and Societal Transitions, 2013, 101: 1-15.
[20] Gartner Hype Cycle [EB/OL]. [2016-07-25]. http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp.

作者贡献声明:

任智军: 提出研究思路, 设计研究方案, 负责 LDA 和相似度实验, 论文起草及最终版本修订;
乔晓东: 提出研究思路, 设计研究方案;
张江涛: 采集、清洗和分析数据, 负责 LDA 实验。

利益冲突声明:

本文作者接受过中国科学技术信息研究所的数据支持。

支撑数据:

支撑数据[1]由作者自存储, E-mail: renzhijun@cnpat.com.cn; 支撑数据[2-3]见期刊网络版 http://www.infotech.ac.cn。
[1] 任智军, 乔晓东, 张江涛. electric_car.txt. 电动汽车整合数据.
[2] 任智军, 张江涛. 电动汽车主题模型.xml. 论文与专利技术主题模型实验结果.
[3] 任智军, 张江涛. Resource.doc. 提供给专家资源列表.

收稿日期: 2016-01-22
收修改稿日期: 2016-04-18

Discover Emerging Technologies with LDA Model

Ren Zhijun[1, 2] Qiao Xiaodong[1] Zhang Jiangtao[2]
[1](Institute of Scientific & Technical Information of China, Beijing 100038, China)
[2](The China Patent Information Center, The State Intellectual Property Office, Beijing 100088, China)

Abstract: [**Objective**] To identify emerging technologies from academic papers and patents. [**Methods**] We adopted the Latent Dirichlet Allocation (LDA) model to find technical topics and used the similarity theory to retrieve emerging technologies from the electric car data. [**Results]** The proposed method was more efficient than exisiting ones. It reduced the subjectivity of the experts' evaluation and the amount of data to be analyzed. [**Limitations**] We did not include the expert scoring experiment in this study, thus, we could not compare the new model's performance with those involving human judgements. [**Conclusions**] The proposed model could identify emerging technologies effectively and then reduce the document reading load of the experts.

Keywords: Emerging technology Paper Patent Electric car Technology similarity